RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2026, **16**, 16613

Industrial wastewater identification based on HPLC combined with data standardization and ensemble learning algorithms

Siyao Li,[a] Haiyan Qin,[b] Xi Wu[a] and Zhirong Suo*[a]

With the continuous development of industry, accidental or unauthorized discharges of wastewater from manufacturing enterprises have increasingly severe impacts on the environment. Rapid and accurate identification of industrial wastewater sources is of great significance for enhancing regulatory oversight and environmental protection. In this study, we propose a novel approach for discriminating the sources of industrial wastewater by integrating data standardization with ensemble learning. High-performance liquid chromatography (HPLC) is employed to collect wastewater samples. To ensure data consistency and accuracy, a local stretching alignment method combined with Gaussian fitting is introduced for precise peak alignment in chromatographic data. We compare the modeling performance of two ensemble learning algorithms: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). To further improve model accuracy, hyperparameter optimization is conducted using the Optuna framework. The models are systematically evaluated through five-fold cross-validation. Experimental results show that the optimized RF model achieves an average cross-validation accuracy of 97.87%, a test set accuracy of 98.28%, and an F1 score of 0.9799, the accuracy on the newly collected samples reached 95.08%, demonstrating excellent overall performance and a well-balanced trade-off between precision and recall. This approach provides an efficient and reliable analytical tool for tracing the sources of industrial wastewater.

Received 4th January 2026
Accepted 19th March 2026

DOI: 10.1039/d6ra00080k

rsc.li/rsc-advances

1 Introduction

With the rapid expansion of industrialization, the number and scale of industrial parks have grown markedly, thereby driving the rapid growth of the national economy. However, this expansion has also led to the discharge of large volumes of toxic wastewater, severely impairing water resources.[1–3] The increasing number of enterprises within these parks has led to the generation of industrial wastewater characterized by high total volume, complex composition, and elevated concentration, resulting in excessively high treatment costs.[4] Consequently, some enterprises illegally discharge industrial wastewater directly without adequate treatment, which substantially affects the performance of wastewater treatment plants, accelerates the deterioration of their structural integrity, increases their operational load, and ultimately leads to system failures.[5] This untreated wastewater is eventually released into natural water bodies, causing ecological damage and posing serious risks to human health.[6] Therefore, the rapid classification and identification of wastewater from different enterprises, as well as the traceability of illegally discharged effluents to their sources, are essential for strengthening the supervision of industrial parks.

Currently, numerous methods exist for identifying water pollution. For instance, traceability methods that rely on conventional water quality indicators, such as chemical oxygen demand (COD), ammonia nitrogen (NH_3^+-N), and total phosphorus (TP), are widely employed.[7,8] While these methods can indicate whether water quality is abnormal, they often provide limited information, making it challenging to accurately distinguish and identify wastewater from different enterprises.[9] Furthermore, analytical methods based on isotopes and trace elements have gained extensive application in river basin pollution source traceability.[10] Sun *et al.*[11] used isotopes $\delta^{13}C$ and $\delta^{15}N$ to trace pollution sources and successfully identified the source of organic matter in the Xinhe Estuary of the Yongding River as originating from urban wastewater and terrestrial sources. However, the elemental composition of industrial wastewater is typically complex and highly similar, and this method cannot effectively distinguish among sources. Therefore, it is imperative to explore other, more efficient identification methods.

To improve identification accuracy, researchers have increasingly been developing methods based on advanced indicators. For example, spectroscopic analytical techniques

[a]*School of Materials and Chemistry, Southwest University of Science and Technology, Mianyang 621010, Sichuan, People's Republic of China. E-mail: suozhirong@163.com*
[b]*College of Life Sciences and Agri-Forest, Southwest University of Science and Technology, Mianyang 621010, Sichuan, People's Republic of China*

such as ultraviolet-visible (UV-Vis),[12] near-infrared (NIR),[13] and fluorescence (FLD)[14,15] are based on the principle that the absorption or emission of a substance at specific wavelengths is governed by its molecular structure.[16] When pollutants produced by different enterprises exhibit similar molecular structures, the discriminatory capability of these spectral analysis methods becomes limited. Additionally, spectral analysis techniques are susceptible to external environmental interference during detection, and the resulting data often require complex processing and analysis methods.[17] High-performance liquid chromatography (HPLC) has become an important technique for identifying pollutants in water due to its high speed, sensitivity, accuracy, and excellent capability for separating complex samples.[18] For instance, József *et al.*[19] utilized HPLC to identify four painkillers in Danube water, while Gure *et al.*[20] integrated HPLC-diode array detection (HPLC-DAD) with ion pair-assisted liquid–liquid extraction (IPA-LLE) to identify six sulfonylureas and four organophosphorus pesticides from three different environmental water bodies in Ethiopia. Although HPLC can accurately and sensitively detect trace pollutants in effluent environments, due to the complex composition of industrial wastewater, it is difficult to achieve rapid identification solely with this method. Machine learning algorithms possess strong logical computing capabilities and demonstrate excellent performance in analyzing complex, nonlinear, and multidimensional data.[21] Lu *et al.*[22] analyzed the relative content of eight pigments in olive oil from five major producing regions in China using HPLC and combined three machine learning algorithms, random forest (RF), *k*-nearest neighbor (KNN), and decision tree (DT), to achieve rapid classification of olive oil producing origins, with the RF model achieving a classification accuracy of 96%. Zhong *et al.*[23] established HPLC fingerprinting of the polysaccharides from Citri Reticulatae Pericarpium (Chenpi) and assessed the capabilities of nine machine learning algorithms to discriminate between different varieties of Chenpi. Among these, five models: linear discriminant analysis (LDA), support vector machine (SVM), artificial neural network (ANN), logistic regression (LR), and quadratic discriminant analysis (QDA) achieved accuracy, precision, recall, and F1-score values all greater than 0.888. However, to date, no reports have been published on the application of HPLC combined with machine learning algorithms for the identification of industrial wastewater.

In this paper, we present a method for identifying wastewater from four enterprises, Sichuan Dongcai New Material Co., Ltd (DC), Mianyang Heze Chemical Co., Ltd (HZ), Sichuan Jin'an Environmental Technology Co., Ltd (JA), and Sichuan Qisai Microelectronics Co., Ltd (QS), that discharging into a centralized wastewater treatment plant within an industrial park in Mianyang, China. HPLC is employed to capture the characteristic features of the wastewater, followed by a series of preprocessing steps applied to the raw data. The RF and extreme gradient boosting (XGBoost) algorithms are then implemented to identify these industrial wastewaters. By comparing the performance of the two algorithms, the RF model is selected as the more appropriate baseline model for this study. In addition, Savitzky–Golay (SG) smoothing is applied for data preprocessing, and the Optuna algorithm is used to optimize the RF model's parameters. The optimized RF model exhibits a notable improvement in accuracy. This method enables rapid identification of wastewater from different enterprises and provides reliable evidence for supervising enterprises, facilitating rapid identification of wastewater across the industrial park.

2 Data collection

2.1 Samples and reagents

In an industrial park in Mianyang, China, four enterprises, DC, HZ, JA, and QS, that discharge into the same centralized wastewater treatment plant were selected. Sampling was conducted at different wastewater treatment stages within each company, with specific sampling points detailed in Table 1. The sampling period spanned from June to July 2025. Three water samples were collected each day, each with a volume of 500 mL, at 9:30 AM, 12:30 PM, and 3:30 PM. The three samples were then mixed in equal volumes to form a composite sample. JA and QS exhibited intermittent discharge patterns, releasing wastewater only on Mondays, Wednesdays, and Fridays. Sampling was consistently scheduled during periods of normal enterprise operation and active wastewater discharge to ensure that the collected water samples were representative of the actual discharged wastewater.

A total of 290 samples were collected in this study, and the organic glass water sampler (Henan Xinchangyuan Experimental Equipment Co., Ltd, China) was used for on-site sampling. The collected water samples were stored in 500 mL high-borosilicate silicon glass bottles (Chengdu Huangyu Experimental Co., Ltd, China), promptly transported to the laboratory, stored at 4 °C, and filtered through a 0.45 μm membrane filter (Tianjin Jinteng Experimental Equipment Co., Ltd, China). The filtrate was transferred into HPLC vials. HPLC-grade methanol was purchased from Hunan Tengma New Material Co., Ltd (Hunan, China), and ultrapure water was produced using a Milli-Q water purification system.

We calculated the similarity between the chromatographic data obtained at all sampling times. The Pearson correlation coefficient was used to assess the similarity among samples from the four enterprises. The similarity between samples from each enterprise exceeded 93%, indicating good stability in the wastewater composition, which is conducive to the subsequent identification of wastewater from the four enterprises.

2.2 HPLC analysis

2.2.1 Chromatographic method. The analysis was conducted using a Wukong K2025 series HPLC system (Wukong Scientific Instrument Co., Ltd, Shanghai, China) equipped with a quaternary pump, an autosampler, a column oven, and a diode-array detector (DAD). Chromatographic separation was achieved using a Venusil XBP C18 column (4.6 mm × 250 mm, 5 μm). The mobile phase consisted of methanol (A) and ultrapure water (B). The gradient elution program was 0–8 min, 10–

Table 1 Sampling points for the four enterprises

Enterprises	Sampling point
DC	Category I wastewater inlet, category II wastewater inlet, fine screen chamber, clear water tank, PV resin production wastewater collection tank, cleaning wastewater collection tank, specialty resin production wastewater collection tank, secondary sedimentation tank, wastewater outlet, final discharge outlet of the plant
HZ	General wastewater inlet, primary sedimentation tank, hydrolysis tank, aerobic tank, secondary sedimentation tank, final sedimentation tank, final wastewater discharge outlet
JA	General wastewater inlet, coagulation sedimentation tank, primary sedimentation tank, anoxic tank, aerobic tank, secondary sedimentation tank, final wastewater discharge outlet.
QS	General wastewater inlet, primary PAM dosing tank, primary sedimentation tank, secondary PAM dosing tank, reconditioning tank, final wastewater discharge outlet

100% A; 8–11 min, 100% A; 11–12 min, 100–10% A; 12–17 min, 10% A; the column temperature was maintained at 30 °C, the flow rate was set to 1.0 mL min^{-1}, the injection volume was 10 μL, and the DAD was operated at a wavelength of 230 nm, 250 nm, 280 nm, 310 nm. Data acquisition and processing were performed using WOOKING LAB Lite software.

Data at 230 nm, 250 nm, 280 nm, and 310 nm were examined. The chromatograms at 230 nm, 250 nm, and 280 nm exhibited numerous peaks with identical retention times, resulting in similar overall chromatographic profiles and introducing interference, thereby reducing the model's classification accuracy. In contrast, at 310 nm, the overall chromatographic profiles showed distinct differences, and preliminary classification predictions yielded better results than at 230 nm and 280 nm. Therefore, 310 nm was selected as the study wavelength.

2.2.2 Method validation. The same filtered sample was injected at 0, 3, 6, 9, 12, and 24 h. The RSDs of the retention times of the main chromatographic peaks for the four enterprises ranged from 0.09% to 0.29%, and the RSDs of the peak areas ranged from 1.83% to 2.87%, indicating good stability of the samples within 24 hours. Six consecutive injections of the same sample yielded RSDs in retention times of 0.02–0.15% and in peak areas of 1.25–1.92%, demonstrating good instrument precision. From the same water sample, six portions were filtered to evaluate repeatability. The RSDs of retention times and peak areas for the four enterprises were between 0.08–0.35% and 2.21–2.65%, respectively. The experiments on precision, stability, and repeatability demonstrated that the method was reliable.

3 Method

In this study, HPLC was integrated with ensemble learning algorithms to identify different enterprises' wastewater. We used overall chromatographic data to construct a model to identify wastewater from different enterprises, based on differences in their complete chromatographic data, without considering specific peak identities. The overall research framework is illustrated in Fig. 1. Initially, wastewater samples were analyzed using HPLC, and the resulting data were filtered using negative detection and the Pearson correlation coefficient. The chromatographic peaks were aligned using the local stretching method and Gaussian fitting, while the characteristic scales were standardized through interpolation. Subsequently, the performance of the RF and XGBoost models was compared, leading to the selection of RF as the predictive model for this study. Hyperparameters were optimized using the Optuna algorithm, and the effects of four data preprocessing methods, first derivative (D1st), second derivative (D2nd), baseline correction (BC), and SG smoothing, on the accuracy of the model were evaluated. Finally, the model's performance was assessed using average accuracy and F1-score obtained from five-fold cross-validation.

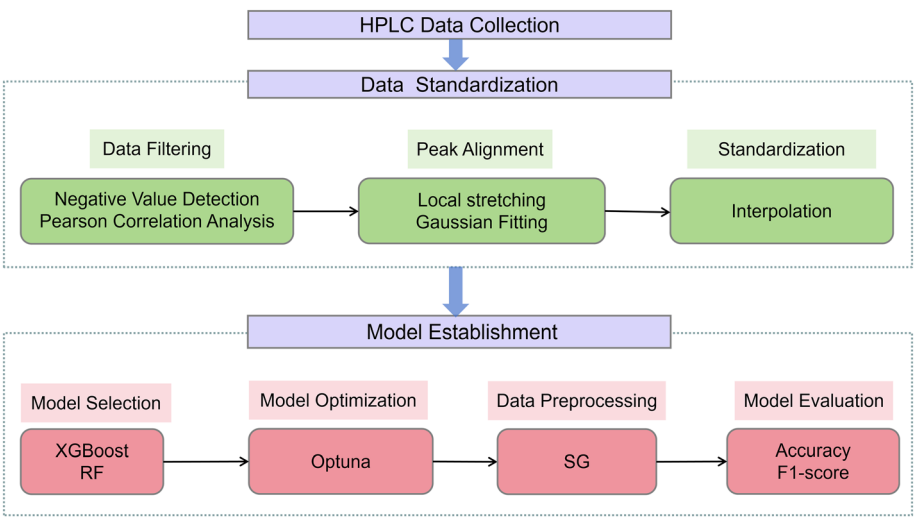

Fig. 1 Over all flow chart.

3.1 Data standardization

Due to differences in instruments and inconsistent experimental procedures, deviations often occur in the number and interval of retention time features in HPLC data. This results in data that cannot be directly used for ensemble learning modeling. Therefore, it is essential to first read and format the water sample data provided by each enterprise to ensure data consistency. Subsequently, abnormal samples were identified through negative value detection and Pearson correlation coefficient analysis, resulting in the generation of abnormal reports for data cleaning and subsequent analysis.

After data cleaning, the scipy.signal.find_peaks function is employed to extract the peaks from the samples. This method identifies significant chromatographic peaks based on specified parameters, such as peak height, adjacent peak spacing, and prominence. Following peak extraction, the retention times of each peak are calculated, and the frequency and distribution of these retention times are analyzed to provide a foundation for subsequent standard retention time setting. Based on the frequency and proportion of retention times, peaks with an occurrence frequency greater than 50% within each enterprise were selected as standard retention times, and the chromatographic peaks were aligned using a local stretching method. Concurrently, the peak area is corrected by Gaussian fitting to ensure consistency across samples at the standard retention time.

The specific procedure is illustrated in Fig. 2, which presents a schematic diagram of peak alignment. First, peak extraction is performed. The extracted peaks are then matched with the standard retention times using a threshold of 0.13 minutes. A match is considered successful if the difference between the extracted peak and the standard retention time is less than 0.13 minutes. A greedy strategy is used to ensure that each standard peak corresponds to a single extracted peak. For successfully matched peaks, local stretching alignment is applied to align the peak center with the standard retention time. Subsequently, Gaussian fits are performed on the aligned peaks to determine peak areas. The parameters of the Gaussian and baseline models (peak height, width, and baseline, with the center fixed to the standard time) are obtained *via* nonlinear least-squares fitting. The peak area is then calculated using trapezoidal numerical integration (scipy.integrate.trapezoid) to obtain the fitted peak area. If the relative error between the fitted and original areas is less than 10%, the peak is considered effectively aligned.

To eliminate scale differences between samples and enhance data comparability, it is also necessary to standardize the chromatographic features, ensuring that the characteristic scales of each sample are aligned. Subsequently, the data are preprocessed using D1st, D2nd, BC, and SG smoothing. The effects of these different preprocessing methods on the model are then compared.

Through these steps, the accuracy and consistency of the data are ensured after removing experimental errors and standardizing the measurements, thereby providing a reliable foundation for subsequent modeling analyses.

3.2 Model establishment

3.2.1 XGBoost. XGBoost is a machine learning algorithm based on gradient boosting that is widely used for classification and regression problems.[24,25] It constructs multiple weak classifiers, typically decision trees, through iterative steps, adjusting the weights of subsequent classifiers based on the errors made by the previous classifiers, thereby enhancing the model's accuracy.[26]

The gradient boosting mechanism employed by XGBoost effectively captures complex patterns in data, particularly in datasets with a large number of features and elevated levels of

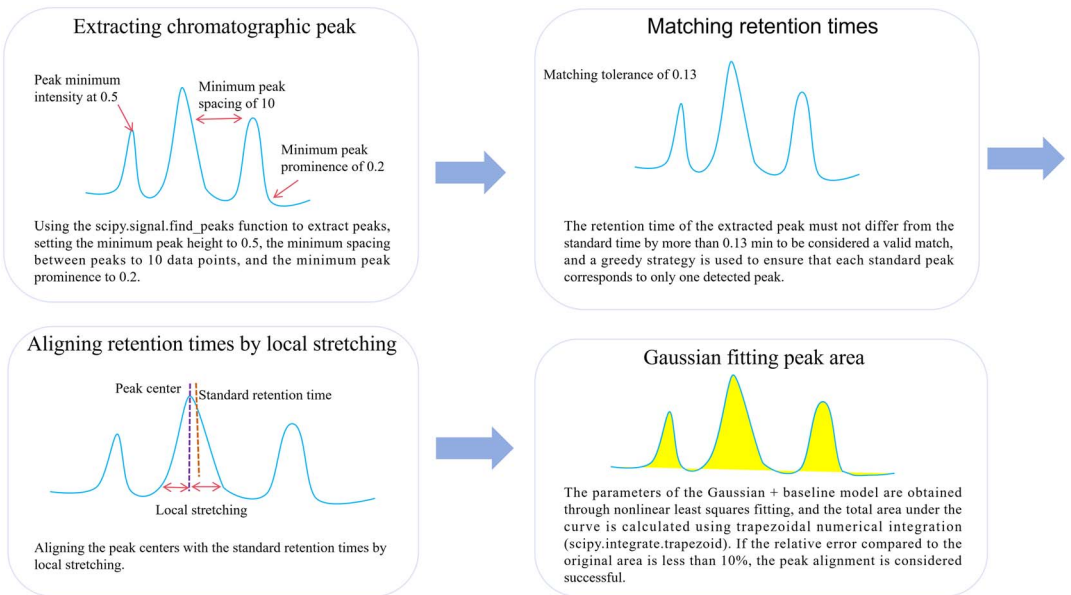

Fig. 2 Schematic diagram of peak alignment.

noise. This capability can significantly enhance classification accuracy. The built-in L1 and L2 regularization mechanisms in XGBoost effectively mitigate overfitting and improve the model's generalization.[27] Furthermore, the model accommodates class imbalance in the data and can optimize classification performance by adjusting class weights.

In the classification task involving data from four enterprises, the initial focus should be on the processing of classification labels. In multi-class classification tasks, these labels serve to distinguish between different categories. The labels are abbreviations for the four enterprises: DC, HZ, JA, and QS.

XGBoost's software implementation requires numeric labels. When using the XGBoost model for a four-class classification problem, the categorical labels DC, HZ, JA, and QS must be converted to numeric labels. This conversion can be achieved through label encoding. Where each categorical label is mapped to a unique integer: DC is encoded as 0, HZ as 1, JA as 2, and QS as 3. It is essential to numerically encode the labels starting from 0, which imposes additional requirements for label processing. Furthermore, XGBoost offers a broader range of hyperparameters, and parameter tuning for multi-class classification problems may require more time and computational resources.

3.2.2 RF. RF is a classification algorithm grounded in ensemble learning principles. It constructs classifiers by aggregating multiple independent decision trees, classifying outcomes using the majority voting principle.[28,29] In multi-classification tasks, RF efficiently handles multiple categories. Compared to XGBoost, its parameter tuning is relatively straightforward, making it suitable for rapid deployment.[30]

Unlike XGBoost, RF does not directly support categorical labels in string format; rather, it processes them indirectly *via* an internal encoding mechanism. This characteristic renders it particularly suitable for rapid classification tasks. Regardless of whether the label is DC, HZ, JA, or QS, the RF model can seamlessly recognize it as a classification label without requiring numerical conversion. This feature enhances the convenience of employing RF for such tasks and facilitates swift modeling. Furthermore, it demonstrates robust tolerance to noise and missing values in the dataset, effectively mitigating the risk of overfitting, especially when dealing with imperfect training data. Because the decision trees in the RF are trained independently, the algorithm benefits from parallel computing, thereby accelerating training and making it well-suited for large-scale datasets.

Although RF can assess feature importance, the overall model integrates multiple decision trees, making it challenging to directly interpret the decision-making process. While RF improves accuracy by increasing the number of trees, merely increasing the number of trees may not yield substantial performance improvements in certain complex classification tasks.

3.3 Classification model evaluation indicators

The evaluation index of the classification model serves as a crucial tool for measuring the model's performance, particularly in assessing its efficacy on the test set. Different evaluation indicators focus on different aspects of model performance. Accuracy and F1-score are commonly used evaluation metrics in classification tasks. Accuracy is the ratio of correctly predicted samples to the total number of samples and serves as an indicator of the model's overall predictive performance. The F1-score is a harmonic mean that combines precision and recall into a single measure of model performance. The formula is as follows.

$$\text{Accuracy} = \frac{\text{TP}_{\text{all}}}{\text{TS}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{F1} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

TP (True Positive): the number of samples correctly predicted as belonging to this category (*i.e.*, the value on the diagonal). TP represents the sum of the number of correctly predicted samples for all categories.

TS (Total Samples): the total number of all samples.

FN (False Negative): the number of samples with true labels in this category but predicted as belonging to another category (*i.e.*, the sum of non-diagonal elements in this row).

FP (False Positive): the number of samples with true labels of other categories but incorrectly predicted as belonging to this category (*i.e.*, the sum of off-diagonal elements in this column).

In this study, model performance is primarily evaluated using two indicators: accuracy and F1-score. These metrics provide crucial quantitative evidence of the model's classification capability, particularly in multi-class classification tasks, where they comprehensively reflect its performance.

3.4 Optimization algorithm

This study mainly uses grid search and the Optuna algorithm to optimize the parameters.

Grid search is a hyperparameter optimization method that identifies the optimal hyperparameter combination by exhaustively evaluating all possible combinations within a specified hyperparameter space.[31] This technique involves establishing a series of predefined candidate values for each hyperparameter of the model (such as learning rate, tree depth, *etc.*), subsequently training all possible combinations of these hyperparameters, evaluating the model's performance, and ultimately selecting the combination that yields the best performance. When the hyperparameter space is limited and computational resources permit, grid search proves to be an effective optimization method. It is particularly suitable for tasks characterized by relatively low computational costs and rapid model training.

Optuna is an efficient framework for automated hyperparameter optimization. Grounded in the principles of Bayesian

optimization, Optuna aims to enhance the performance of machine learning models by intelligently searching for and optimizing their hyperparameters. Its core principle is to use historical experimental results to inform hyperparameter selection and achieve optimal convergence within the hyperparameter space through a systematic optimization process.[32,33] Specifically, Optuna constructs a surrogate model, known as a probability model, based on the Tree-structured Parzen Estimator (TPE) algorithm. It adjusts the hyperparameter search strategy based on feedback from historical experiments. Each time new hyperparameters are explored, Optuna evaluates their impact on model performance and updates the search space accordingly, thereby increasing the search efficiency. Through this approach, the TPE algorithm can accurately estimate the optimal hyperparameter configuration with fewer trials, ultimately converging on the global optimal solution.[34]

3.5 The division of the training set and the test set

In this study, the train_test_split function is used to split the dataset, which supports stratified sampling *via* the stratify parameter. During the partitioning process, 80% of the data is allocated for training, while the remaining 20% is reserved for testing. To ensure experiment repeatability, the random seed is set to random_state = 42, ensuring that the datasets split in each experiment remain consistent and facilitating stable experimental results. Weights are assigned based on the number of samples in various categories, which enhances the model's attention to underrepresented categories and mitigates the adverse effects of class imbalance on model training. The specific quantities of the training and test sets are presented in Table 2.

To enhance the model's generalization and ensure stability across datasets, five-fold cross-validation is used as the primary evaluation method. The five-fold cross-validation divides the training set into five subsets and performs the following steps:

(1) Data partitioning: first, the dataset is randomly divided into five non-overlapping subsets. Each subset is of equal size and preserves the class distribution of the original dataset, thereby mitigating the potential effects of class imbalance.

(2) Training and verification: in each round of training, four subsets are selected to form the training set, while the remaining subset is designated as the verification set. Consequently, the model is trained and evaluated five times, using different training and validation sets each time.

(3) Performance evaluation: after each training round, the performance indicators (such as accuracy, F1-score, *etc.*) of the model on the validation set are calculated. The overall performance of the model is represented by the average results of these five evaluations, which ensures the stability and reliability of the model assessment.

The five-fold cross-validation method effectively mitigates the risk of model overfitting. Each sample is evaluated across multiple training and validation sets, thereby enhancing the model's generalization. The model's overall performance is derived by averaging the evaluation results across folds, making it a more reliable approach than a single training-test set partition.

Table 2 Number division of the training set and the test set

Enterprises	Total sample number	Training set	Testing set	Weight
DC	99	79	20	1/79
HZ	68	54	14	1/54
JA	55	44	11	1/44
QS	68	55	13	1/55

4 Result and discussion

4.1 Data standardization

A total of 290 water samples from four enterprises within the same industrial park were analyzed by HPLC. In HPLC data analysis, the retention time shift of chromatographic peaks often occurs due to factors such as injection differences or instrument state fluctuations. Therefore, standardization is necessary. The significance of standardization in machine learning modeling is that it reduces the impact of factors such as retention time offset and instrument fluctuations, and improves the comparability and stability of data. By standardizing operations, it is possible to ensure that chromatographic data is compared at the same scale, thereby improving the accuracy and generalization ability of the model. Therefore, this study proposes an HPLC peak standardization method that combines anomaly screening with chromatographic peak alignment and Gaussian fitting. Based on the frequency and proportion of retention times in each enterprise, peaks with an occurrence frequency exceeding 50% were selected as standard retention times. The original and standardized HPLC chromatograms of the samples collected from the four enterprises are shown in Fig. 3(a). The figure displays five standard retention times: 2.49 min, 2.98 min, 6.64 min, 13.29 min, and 15.09 min. The peaks at 2.98 min and 15.09 min are prevalent across all four companies, while the peak at 2.49 min occurs only in JA samples, the peak at 6.64 min occurs only in HZ samples, and the peak at 13.29 min occurs only in QS samples. It can be observed that deviations exist between the original retention times of the four enterprises and the standard retention times, and the data scales of the four companies also differ. To correct these deviations, the peak and its region are aligned using local stretching and Gaussian fitting. Finally, the chromatographic data is standardized using interpolation methods to ensure consistency in length and interval, thereby improving the comparability and reliability of subsequent analysis. The HPLC spectra after standardization are shown in Fig. 3(b). It can be seen that the standardized spectra effectively reduce the interference caused by experimental process deviations on the data, which can provide reliable guarantees for the subsequent input data based on machine learning models.

4.2 Model selection

Perform modeling analysis on the full-profile chromatographic data of 290 samples from four enterprises within the same

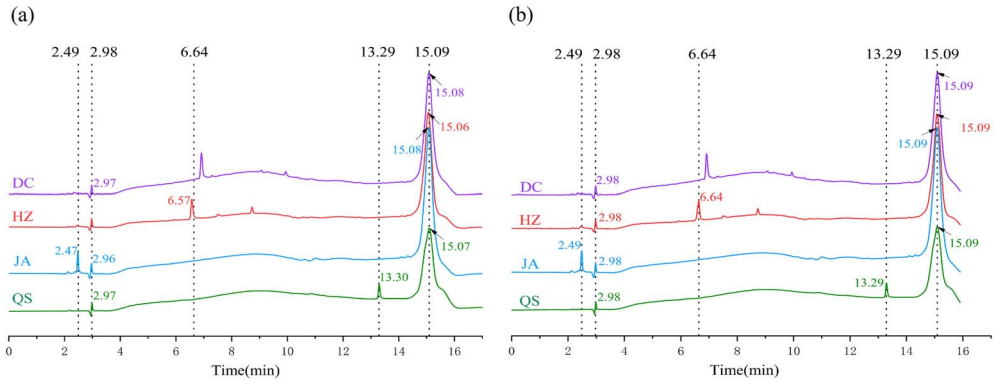

Fig. 3 (a) Raw chromatograms of the four companies, (b) standardized chromatograms of the four companies.

industrial park, and evaluate the results of the two models. We can comprehensively evaluate and select the appropriate model. A comparison of the RF and XGBoost models found that their average accuracy was 82.76%. However, the RF model has its own classifier, capable of assigning classification labels to character classes. The XGBoost model can only recognize numerical labels; therefore, it is necessary to encode categorical variables as numbers to ensure that the model correctly outputs the appropriate labels. The running speed of the two models is compared. To ensure a fair comparison and maintain consistency in common hyperparameters between the RF and XGBoost models (such as the number and depth of trees), the XGBoost model is recorded at 39.67 seconds, whereas the RF model is recorded at 10.67 seconds. During execution, steps such as data preprocessing and model tuning incur significant computational overhead. The RF model includes a built-in classifier and has a short training time. In this study, the RF model is selected for training, as it effectively reduces computational costs and improves system response efficiency while maintaining model performance.

4.3 Model optimization

Initially, the modeling parameters of the original dataset were optimized, and the training set was used for five-fold cross-validation. The objective was to achieve the highest average accuracy in this five-fold cross-validation, guiding model tuning and method selection.

Using the Optuna algorithm, we obtained the optimal values for the model's key hyperparameters: "n_estimators = 469", "max_depth = 40", "min_samples_split = 1", "min_samples_leaf = 1", and "max_features = sqrt". The n_estimators reduces model variance by increasing the number of trees. The max_depth parameter limits the maximum depth of each tree to prevent the model from becoming overly complex and experiencing overfitting. The min_samples_split prevents over-splitting by increasing the minimum number of samples per node. The min_samples_leaf further enhances generalization by increasing the minimum number of samples per leaf node. And max_features increases the model's diversity by limiting the number of selected features per tree, thereby reducing computational costs while suppressing overfitting. These

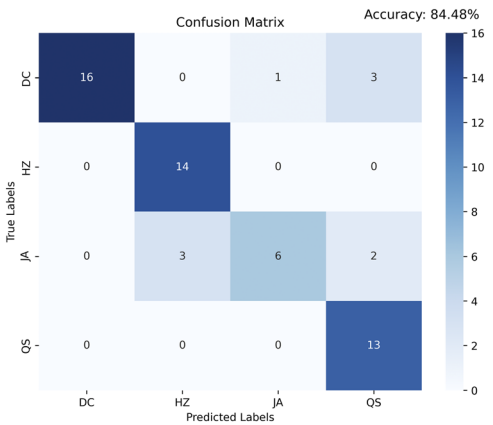

Fig. 4 Confusion matrix of the test set after RF hyperparameter optimization.

hyperparameters work together to control model complexity and improve generalization ability.

After importing the optimal parameters, the average accuracy achieved through five-fold cross-validation is 86.19%. In addition, as shown in Fig. 4, which presents the optimized confusion matrix, the test set accuracy increased from 82.76% to 84.48%.

Continue studying the selection of preprocessing methods with optimized parameters and compare them. The methods include D1st, D2nd, BC, and SG smoothing. Using the average

Table 3 Results of different data preprocessing method

Method	$Accuracy_{cv}$ [a]	$F1_{cv}$ [b]	$Accuracy_P$ [c]	$F1_P$ [d]
Original	86.20%	0.8548	84.48%	0.8244
SG	95.29%	0.9490	96.55%	0.9594
BC	85.36%	0.8492	75.86%	0.7468
D2nd	57.30%	0.5391	68.97%	0.6523
D1st	53.07%	0.4855	60.34%	0.5887

[a] $Accuracy_{cv}$ denotes the average accuracy of five-fold cross-validation. [b] $F1_{cv}$ denotes the average F1 score of 5-fold cross-validation. [c] $Accuracy_P$ denotes the accuracy of the test set. [d] $F1_P$ denotes the F1 score of the test set.

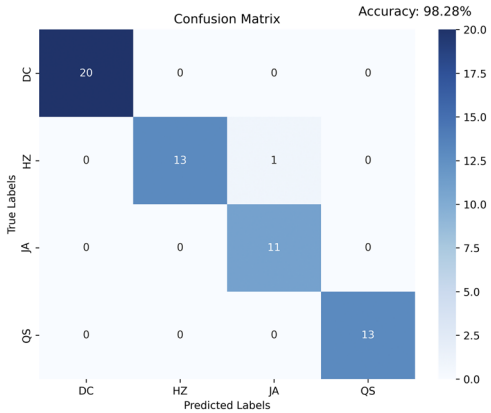

Fig. 5 Confusion matrix of the test set after SG smoothing.

accuracy of five-fold cross-validation as the standard, organize the results of different data preprocessing methods, as shown in Table 3.

According to Table 2, the SG smoothing method achieved the greatest improvement in accuracy during five-fold cross-validation. By using Optuna to optimize the window size and polynomial order for SG smoothing, and using a grid search to enumerate all parameter settings, the optimal configuration was identified: a window size of 11 and a polynomial order of 3. Following this optimization, the average accuracy in five-fold cross-validation was 97.85%, suggesting that the model remains stable across different data subsets. Upon training the model on the complete training set, Fig. 5 shows the identification results of the test set after SG smoothing. It achieved 98.28% accuracy on the test set, demonstrating outstanding predictive performance. The accuracy of both the internal five-fold cross-validation and the external test set validation exceeded 97%, indicating that the model did not exhibit overfitting. Furthermore, the model's F1-score was 0.9799, indicating a commendable balance between precision and recall and demonstrating its effectiveness in identifying samples.

In March 2026, two newly collected batches of samples from four enterprises were used to validate the model's performance. The HPLC data from the newly collected samples were preprocessed using standardization and SG smoothing, without the need to recalculate standard retention times. The RF model achieved an overall accuracy of 95.08% and an F1 score of 0.9475, indicating excellent model performance in identifying industrial wastewater from different companies.

Pairwise cross-mixing of the final wastewater discharge outlet samples from the newly collected first batch of four enterprises was conducted at a 1 : 1 volume ratio to investigate the impact of mixed samples on the model. The identification results for different mixing methods are presented in Table 4. The overall matching rate for the four enterprises was 1. Among the mixed samples, the matching degree for QS was relatively low, indicating that the overall identification performance for QS was not ideal. Therefore, the RF classification model currently has limitations in identifying mixed samples.

Table 4 Identification results of different mixed samples for the four companies (total matching rate of the four companies = 1)

Mixed samples (1 : 1, v/v)	DC	HZ	JA	QS
DC : HZ	0.2768	0.5071	0.1121	0.1040
DC : JA	0.5636	0.0684	0.3382	0.0622
DC : QS	0.5694	0.0501	0.3447	0.0358
HZ : JA	0.2580	0.4234	0.2750	0.0508
HZ : QS	0.1531	0.5305	0.2538	0.0626
JA : QS	0.1940	0.0369	0.5695	0.1996

5 Conclusion

In this study, an HPLC system was employed to collect wastewater samples from four distinct industrial sources. Prior to machine learning modeling, a series of preprocessing steps was applied to enhance data quality and ensure model validity. Obvious outliers were first removed, followed by chromatographic peak alignment using local stretching and Gaussian fitting, and subsequent data standardization to correct experimental deviations. Among the evaluated algorithms, the RF model outperformed XGBoost in classification accuracy and stability after hyperparameter optimization using Optuna and SG smoothing. The final RF model achieved an average five-fold cross-validation accuracy of 97.87 percent, a test set accuracy of 98.28 percent, and an F1 score of 0.9799, indicating a strong balance between precision and recall. The overall accuracy of the newly collected two batches of samples from four enterprises is 95.08%, demonstrating the model's strong identification performance for different industrial wastewaters.

Despite these encouraging results, opportunities for further improvement remain. The model exhibited unsatisfactory identification performance for QS in mixed samples, revealing certain limitations. The current dataset is relatively small and exhibits a slight class imbalance. Future work will focus on expanding the sample pool, diversifying the range of industrial categories, and conducting research on cross-source mixed samples to enhance the model's generalization. Additionally, the framework will be extended beyond classification to quantitative analysis by leveraging chromatographic peak areas, thereby broadening its practical utility. The success of this study, achieved through basic preprocessing and parameter tuning alone, highlights the substantial potential of machine learning in chromatographic water sample analysis. Further gains are expected by integrating advanced preprocessing pipelines, feature engineering strategies, and ensemble techniques to handle increasingly complex real-world datasets.

Author contributions

Siyao Li: investigation, methodology and writing – original draft preparation. Haiyan Qin: investigation, methodology and writing – original draft preparation. Xi Wu: methodology, software and data curation. Zhirong Suo: funding acquisition and supervision.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article is provided in an Excel file.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d6ra00080k.

Acknowledgements

This work is supported by funds from the Industrial Wastewater Biotoxicity Assessment and Treatment Project (25zh0214).

Notes and references

1 J. Lu, Q. Zhou, W. Qi, S. Qu and J. Bi, *Sci. Total Environ.*, 2023, **896**, 165279.
2 Y. Qin, H. Zhang, H. Zhao, D. Li, Y. Duan and Z. Han, *Front. Environ. Sci.*, 2022, **10**, 1004343.
3 L. Zhu, Z. J. B. M. Husny, N. A. Samsudin, H. Xu and C. Han, *Urban Clim.*, 2023, **49**, 101486.
4 D. J. Dürrenmatt and W. Gujer, Identification of industrial wastewater by clustering wastewater treatment plant influent ultraviolet visible spectra, *Water Sci. Technol.*, 2011, **63**, 1153–1159.
5 R. M. Salem, M. S. Saraya and A. Ali-Eldin, *IEEE Access*, 2022, **10**, 6528–6540.
6 L. Lin, H. Yang and X. Xu, *Front. Environ. Sci.*, 2022, **10**, 880246.
7 K. Zhang, S. Wang, S. Liu, K. Liu, J. Yan and X. Li, *Sustainability*, 2022, **14**, 9219.
8 X. Li, C. Li, X. Wang, Q. Liu, Y. Yi and X. Zhang, *Water*, 2022, **14**, 730.
9 P. Yan, X. Zhang, X. Kan, H. Zhang, R. Qi and Q. Huang, *Water*, 2023, **15**, 701.
10 C. Niu, T. Zhai, Q. Zhang, H. Wang and L. Xiao, *Int. J. Environ. Res. Public Health*, 2021, **18**, 11805.
11 C. Sun, Q. Wei, L. Ma, L. Li, G. Wu and L. Pan, *Mar. Pollut. Bull.*, 2017, **115**, 451–458.
12 Z. Wei, Y. Ji, H. Fang, L. Yu and D. Dong, *Water*, 2025, **17**, 790.
13 Z. Xu, X. Li, W. Cheng, G. Zhao, L. Tang, Y. Yang, Y. Wu, P. Zhang and Q. Wang, *Spectrochim. Acta, Part A*, 2023, **302**, 123007.
14 B. Liu, J. Wu, C. Cheng, J. Tang, M. F. S. Khan and J. Shen, *Chemosphere*, 2019, **216**, 617–623.
15 Y. Zhang, X. Liang, Z. Wang and L. Xu, *Sci. Rep.*, 2015, **5**, 16079.
16 W. Cai, C. Ye, F. Ao, Z. Xu and W. Chu, *Water Res.*, 2025, **277**, 123281.
17 Q. Yu, H. Yin, K. Wang, H. Dong and D. Hou, *Water*, 2018, **10**, 1566.
18 P. Gątarek, A. Rosiak and J. Kałużna-Czaplińska, *Crit. Rev. Anal. Chem.*, 2025, **55**, 840–857.
19 T. József, S. R. Kiss, F. Muzslay, O. Máté, G. P. Stromájer and T. Stromájer-Rácz, *Water*, 2023, **15**, 1755.
20 A. Gure, N. Megersa and N. Retta, *Anal. Methods*, 2014, **6**, 4633–4642.
21 Y. Shu, F. Kong, Y. He, L. Chen, H. Liu, F. Zan, X. Lu, T. Wu, D. Si, J. Mao and X. Wu, *Water Res.*, 2025, **268**, 122618.
22 C. H. Lu, Y. Gao, H. Y. Lu, W. J. Shen, J. Muhire, Z. B. Lu, Q. Jing, X. Y. Huang, D. Pei and D. L. Di, *J. Am. Oil Chem. Soc.*, 2025, **102**, 1029–1038.
23 M. Y. Zhong, M. N. Li, W. S. Zou, S. Q. Hu, J. N. Luo, Q. X. Jiang, Q. F. Cao, L. F. Lin, Z. X. Wang, H. Li and W. W. Deng, *Food Chem.*, 2025, **473**, 143053.
24 J. R. Quinlan, *Mach. Learn.*, 1986, **1**, 81–106.
25 X. Li, Y. Jia, D. Zhang, J. Yang and Z. Chen, *Water Int.*, 2023, **48**, 309–321.
26 G. Grekousis, *J. Geogr. Syst.*, 2025, **27**, 169–195.
27 J. Wu, D. Ma and W. Wang, *Water Resour. Plann. Manage.*, 2022, **148**, 04021107.
28 L. Grbčić, I. Lučin, L. Kranjčević and S. Družeta, *J. Hydroinf.*, 2020, **22**, 1521–1535.
29 H. Wei, H. Qiu, J. Liu, W. Li, C. Zhao and H. Xu, *Ecotoxicol. Environ. Saf.*, 2025, **289**, 117499.
30 K. S. More and C. Wolkersdorfer, Predicting and forecasting mine water parameters using a hybrid intelligent system, *Water Resour. Manage.*, 2022, **36**, 2813–2826.
31 H. Chen, Z. Zhang, W. Yin, C. Zhao, F. Wang and Y. Li, *Measurement*, 2022, **189**, 110660.
32 B. A. Dada, N. I. Nwulu and S. O. Olukanmi, *Smart Agric. Technol.*, 2025, **12**, 101136.
33 X. Xiao, Y. Zou, J. Huang, X. Luo, L. Yang, M. Li, P. Yang, X. Ji and Y. Li, *Geomat. Nat. Hazards Risk*, 2024, **15**, 2347421.
34 Y. Zhou, Z. Dong and X. Bao, *Appl. Sci.*, 2024, **14**, 3719.